An Integrated Segmentation and Classification Approach Applied to Multiple Sclerosis Analysis

Ayelet Akselrod-Ballin[1], Meirav Galun[1], Moshe John Gomori[2]
Massimo Filippi[3], Paola Valsasina[3], Ronen Basri[1], Achi Brandt[1]
[1]Dept. of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot, Israel
[2] Dept. of Radiology, Hadassah University Hospital, Jerusalem, Israel
[3]Neuroimaging Research Unit, Hospital San Raffaele, Milan, Italy

Abstract

We introduce a novel multiscale approach that combines segmentation with classification to detect abnormal brain structures in medical imagery, and demonstrate its utility in detecting Multiple Sclerosis lesions in 3D multi-channel MRI data. Our method uses segmentation to obtain a hierarchical decomposition of a multi-channel, anisotropic MRI scan. It then produces a rich set of features describing the segments in terms of intensity, shape, location, neighborhood relations, and anatomical context. These features are then fed into a decision tree-based classifier, trained with data labeled by experts, enabling the detection of lesions in all scales. Unlike common approaches that use voxel-by-voxel analysis, our system can utilize regional properties that are often important for characterizing abnormal brain structures. We provide experiments showing successful detections of lesions in two types of real MR images.

I. Introduction

Identifying 3D brain structures in medical imagery, particularly in Magnetic Resonance Imaging (MRI) scans, is important for early detection of tumors, lesions, and abnormalities, with applications in diagnosis, follow-up, and image-guided surgery. Computer aided analysis can assist in identifying brain structures, extract quantitative and qualitative properties of structures, and evaluate their progress over time. In this paper we present a novel method for detecting abnormal brain structures focusing on 3D MRI brain data containing scans of Multiple Sclerosis (MS) patients.

Manual or interactive segmentation by human experts is time-consuming, expensive, and suffers from considerable inter- and intra- rater variability. In addition it is difficult for the human expert to combine information from several slices and multiple channels when multi spectral MRI data is examined. While semi automatic methods ([1],[2],[3]) significantly improve the inter- and intra- rater variability they still depend on varying degrees of human intervention, which often are not as robust, reproducible and reliable as the analysis that would be made by top expert radiologists. Consequently an automatic quantitative analysis of MS in MRI has become increasingly important. Hence, many automatic algorithms for MS segmentation have been published in literature. Zijdenbos et al. [4] developed an automatic pipeline for T1- T2- and PD-weighted images based on a supervised artificial neural network (ANN) classifier and validated it extensively on multi-center data. Statistical models have been widely employed ([5],[6],[7]). Wells et al. [8], introduced the expectation maximization (EM) segmentation for MRI which simultaneously estimated the bias field correction and classifies intensity regions in the brain using a Gaussian distribution to model the tissue class intensities. Warfield et al. ([5],[9]) combined elastic atlas registration with a statistical approach based on the EM classifier. Where a non-linear registration process matched between a digital brain atlas to the patients brain. In [6] three pipelines reproducibility and accuracy were compared for quantitative white matter(WM) signal abnormalities. The pipeline that combined all components, EM, template driven segmentation(TDS) employing a deformable digital anatomical atlas and an heuristic connectivity based PVE correction component, demonstrated the highest accuracy. In a recent work [10], the authors expanded this system to an automated three-channel MRI segmentation pipeline for MS lesions subtypes, and showed that it improved the sensitivity, specificity and accuracy. Van Leemput et al. [7] modeled the intensity of brain tissues by Gaussian Mixture Model (GMM) utilizing the EM framework and extended the algorithm using probabilistic brain atlas maps for initialization of the prior and also for geometric constraint. The MS lesions were detected as outliers with respect to the normal brain model. The scheme applied two constraints. An intensity constraint, required that MS lesions voxel would be brighter than the estimated mean intensity of grey matter (GM) and a contextual constraint required that most of the voxels 3D neighboring voxels would belong to WM or MS lesions.

Another category of studies (e.g. [11], [12], [13], [14]) investigated the time domain for MS lesion segmentation. These studies identify regions as MS lesion based on their temporal evolution profile. A critical concept of these approaches is that lesions, tumors or anatomical structures vary over time either due to the pathological process or under the effect of therapy. This may cause difficulties in cases where limited temporal information of the disease is available for the subject.

Automatic detection of abnormal brain structures, and particularly MS lesions, is difficult. Abnormal structures exhibit extreme variability. Their shapes are deformable, their location across patients may differ significantly, and their intensity and texture characteristics may vary. Detection techniques based on template matching [15] or more recent techniques based on constellations of appearance features (e.g., [16]), which are common in computer vision, are not well suited to handle such amorphous structures. Consequently most medical applications commonly approach this problem by applying classification algorithms that rely on a voxel-by-voxel analysis, where the image intensity and atlas probability values are the features of interest (e.g., [5], [6], [8], [4]). There are a few exceptions (e.g., [7], [14]) of approaches that do employ the voxel intensity of neighbors, for example by using a Markov Random Field (MRF) model in the contextual constraint (see [7]). However none of the approaches utilize regional statistical properties at different scales, particularly properties related to the shape, boundaries, and texture statistics.

This paper introduces a novel multiscale approach that combines segmentation with classification to detect abnormal 3D brain structures. Our method is based on a combination of a powerful multiscale segmentation algorithm, Segmentation by Weighted Aggregation (SWA) ([17], [18], [19]), a rich feature vocabulary describing the segments, and a decision tree-based classification of the segments.

The methodology presented here differs from published MS analysis works in several aspects. First, to the best of our knowledge we are the first to present a **multi-scale approach** to the segmentation of MS by applying a graphical model. We adapt the SWA algorithm to handle 3D multi-channel MRI scans and anisotropic voxel resolutions. These allow the algorithm to handle realistic MRI scans. Second, we are not aware of an approach that uses such a rich set of multiscale **features** to characterize MS lesions. The bank of features we use characterize each aggregate in terms of intensity, texture, shape, and location. These features were selected in consultation with expert radiologists. All the features are computed as part of the segmentation process, and they are used in turn to further affect the segmentation process. The classification step examines each aggregate and labels it as either lesion or non-lesion. This classification is integrated across scale to determine the voxel classification of the lesions. Three, we use an automatic learning process. So that adaptation to new types of task and data can be done via training. In contrast with other MRI segmentation approaches that are usually tailored for the task dealt with. We suggest **a system that integrates segmentation and learning and is general and flexible**. Hence, we have reported results with parts of this framework for brain tissue segmentation and rat uterus delineation in MRI ([34], citeAkselrod-BallinSPIE:2006). Forth like other approaches the method is fully automatic due to the use of a probabilistic brain atlas. Moreover the approach includes spatial information of the **cerebellum** due to the difficulty in detection of MS in this area. Fifth, and additional benefit of the algorithm is its ability to provide a soft classification rather than just a binary result. Since it is well known that the ’ground truth’ of the lesions may vary among different experts. This property may be useful in clinical trials, to obtain a consistent reproducible full range of results.

By combining segmentation and classification we are able to utilize integrative, regional properties that provide regional statistics of segments, characterize their overall shapes, and localize their boundaries. At the same time, the rich hierarchical decomposition produced by the SWA algorithm allows us to a great extent to circumvent inaccuracies due to the segmentation process. Even when a lesion is not segmented properly, namely when it is not fully covered by one aggregate, we can generally expect to find some aggregate in the hierarchy that sufficiently overlaps it to allow classification.

The utility of the method is demonstrated through experiments on two types of real MRI data showing detection of MS lesions. One multi-channel set that consist of T1-, T2-, and PD-weighted MR images and another single channel data set that consists of FLAIR images, where we are unaware of an automatic MS segmentation study that has been extensively validated with FLAIR data. We compare our approach to the state-of-the-art automated methods evaluating sensitivity, specificity, accuracy, and spatial correspondence between automated segmentations and manual lesion tracings derived from kappa statistics.

The paper is organized as follows. Section II presents the segmentation procedure, the feature extraction method and the classification model in our system. In section III results on two types of real MRI data are presented. Section IV follows with a discussion and conclusions.

An earlier version of this work has appeared in an IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) [20].

II. Methods: Integrated System

This section describes our system for detecting abnormal brain structures. In a training phase our system obtains as input several MR scans along with a delineation of the lesions in these scans. The system uses segmentation to provide a complete hierarchical decomposition of the 3D data into regions corresponding to both meaningful anatomical structures and lesions. Each aggregate is equipped with a collection of multiscale features. Finally, a classifier is trained to distinguish between aggregates that correspond to lesions from those that correspond to non-lesions. The block diagram in figure 1 presents the integrated systems modules.

Once the classifier is trained we proceed to apply our approach to an unlabeled test data. At this stage the system obtains as input an MRI scan of a single brain. It then segments the scan and extracts features to describe the aggregates. Finally, each aggregate is classified as either a lesion or a non-lesion, and the voxel occupancy of the lesions is determined.

A. Segmentation

We use the Segmentation by Weighted Aggregation (SWA) algorithm [17], [18], [19], which we extend to handle 3D multi-channel and anisotropic data. In this section we review the SWA algorithm along with our extensions, more details on the general segmentation model and its motivations can be found in ([17], [18], [19]).

*1) **Segmentation Framework**:* Given a 3D MRI scan, a 6-connected graph $G = (V, W)$ is constructed as follows. Each voxel i is represented by a graph node i, so $V = \{1, 2, \ldots, N\}$ where N is the number of voxels. A weight is associated with each pair of neighboring voxels i and j. The weight w_{ij} reflects the contrast between the two neighboring voxels i and j

$$\omega_{ij} = e^{-\alpha|I_i - I_j|} \tag{1}$$

where I_i and I_j denote the intensities of the two neighboring voxels, and α is a positive pre-defined constant that is determined with experience ($\alpha = 15$ in our MRI experiments). [*meirav - The rational for this function form] We define the saliency of a segment by applying a normalized-cut-like measure as follows. Every segment $S \subseteq V$ is associated with a state vector $u = (u_1, u_2, \ldots, u_N)$, representing the assignments of voxels to a segment S

$$u_i = \begin{cases} 1 & \text{if } \ i \in S \\ 0 & \text{if } \ i \notin S. \end{cases} \tag{2}$$

The **saliency** Γ associated with S is defined by

$$\Gamma(S) = \frac{u^T L u}{\frac{1}{2} u^T W u}, \tag{3}$$

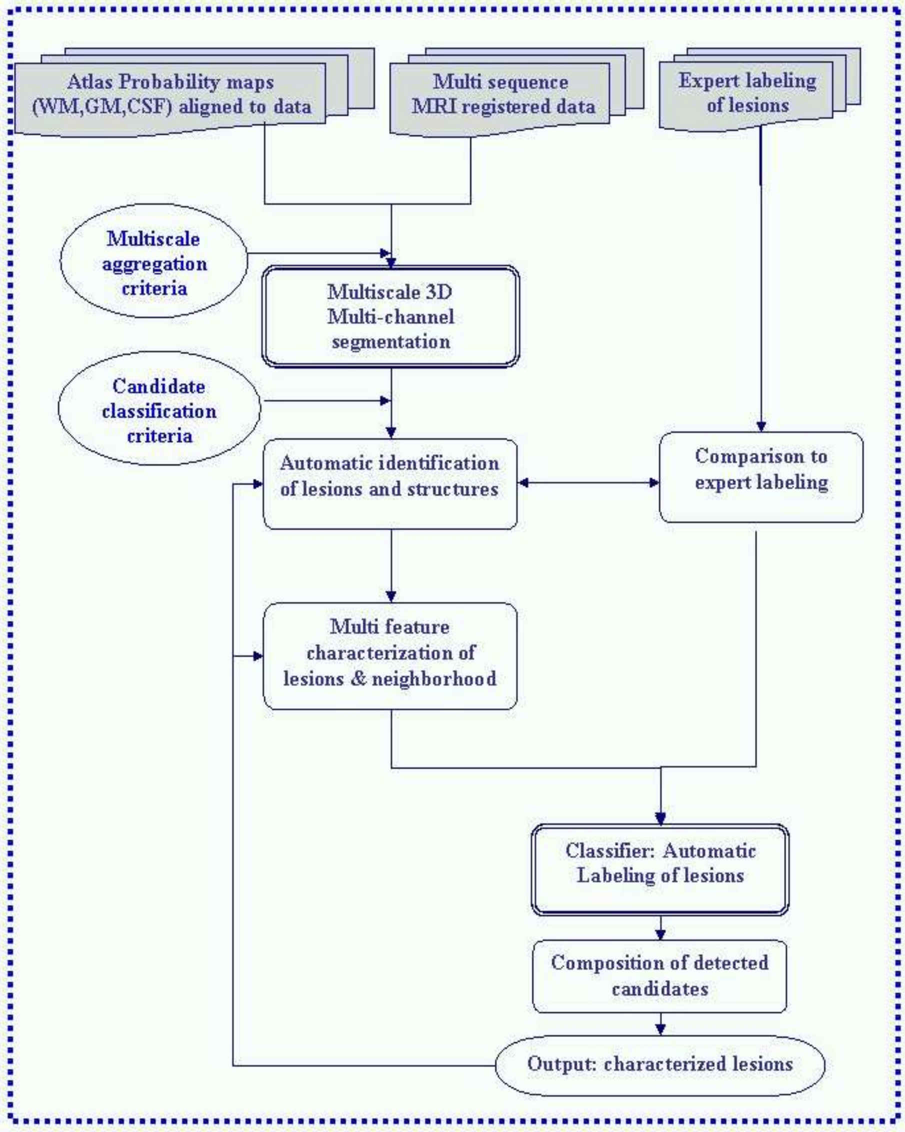

Atlas Probability maps (WM,GM,CSF) aligned to data
Multi sequence MRI registered data
Expert labeling of lesions
Multiscale aggregation criteria
Multiscale 3D Multi-channel segmentation
Candidate classification criteria
Automatic identification of lesions and structures
Comparison to expert labeling
Multi feature characterization of lesions & neighborhood
Classifier: Automatic Labeling of lesions
Composition of detected candidates
Output: characterized lesions

Fig. 1. Outline of the entire integrated system.

where the matrix W includes the weights w_{ij}, and L is the Laplacian matrix of G whose elements are

$$l_{ij} = \begin{cases} \sum_{k\,(k \neq i)} w_{ik} & i = j \\ -w_{ij} & i \neq j. \end{cases} \tag{4}$$

The saliency measure sums the weights along the boundaries of S normalized by the internal weights. Segments that yield small values of $\Gamma(S)$ are considered salient. If we allow arbitrary real assignments to u, the minimum for Γ is obtained by the minimal generalized eigenvector u of $Lu = \lambda W u$, with the condition that $\lambda > 0$. This equation is in fact equivalent to the normalized cuts solution [21].

Our objective is to find those partitions characterized by small values of Γ. To find the minimal cuts in the graph we construct a coarse version of this graph. This coarse version is constructed so that we can use salient segments in the coarse graph to predict salient segments in the fine graph using only local calculations. This coarsening process is repeated recursively,

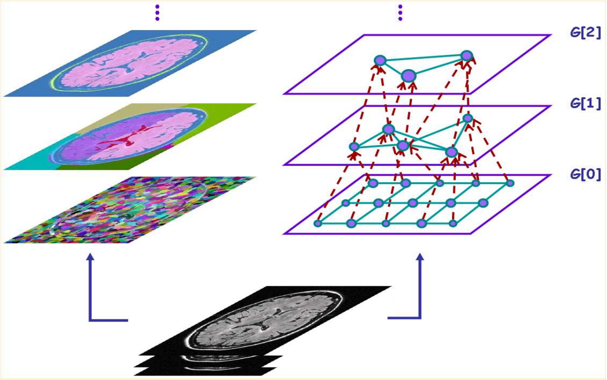


Fig. 2. Illustration of the segmentation pyramid hierarchy. The image presents three graph levels above a 3D MRI input data.

constructing a full pyramid of segments (see Fig. 2). Each node at a certain scale represents an **aggregate** which is a weighted collection of voxels. Each **segment** S, which is a salient aggregate (i.e., $\Gamma(S)$ is low), emerges as a single node at a certain scale.

The coarsening procedure proceeds recursively as follows. Starting from the given graph $G^{[0]} \overset{def}{=} G$, we create a sequence of graphs $G^{[1]}, \ldots, G^{[k]}$ of decreasing size (Fig. 2). As in the general Algebraic Multigrid (AMG) setting [22], the construction of a coarse graph from a fine one is divided into three stages: first a subset of the fine nodes is chosen to serve as the **seeds** of the aggregates (the latter being the nodes of the coarse graph). Then, the rules for interpolation are determined, establishing the fraction of each non-seed node belonging to each aggregate. Finally, the weights of the edges between the coarse nodes are calculated.

Coarse seeds: The construction of the set of seeds C, and its complement denoted by F, is guided by the principle that each F-node should be ”strongly coupled” to C. To achieve this objective we start with an empty set C, hence $F = V$, and sequentially (according to decreasing aggregate volume defined in Sec. II-B) transfer nodes from F to C until all the remaining $i \in F$ satisfy $\sum_{j \in C} w_{ij} \geq \eta \sum_{j \in V} w_{ij}$, where η is a parameter (in our experiments $\eta = 0.2$).

The coarse problem: We define for each node $i \in F$ a coarse **neighborhood** $N_i = \{j \in C, w_{ij} > 0\}$. Let $I(j)$ be the index in the coarse graph of the node that represents the aggregate around a seed whose index at the fine scale is j. An **interpolation** matrix P (of size $N \times n$, where $n = |C|$) is defined by

$$P_{iI(j)} = \begin{cases} \frac{w_{ij}}{\sum_{k \in N_i} w_{ik}} & \text{for } i \in F, j \in N_i \\ 1 & \text{for } i \in C,\, j = i \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This matrix satisfies $u \approx PU$, where $U = (U_1, U_2, ..., U_n)$ is the coarse level state vector. P_{iI} represents the likelihood that an aggregate i at a fine level belongs to an aggregate I at a coarser level. Finally, an edge connecting two coarse aggregates

TABLE I
OUTLINE OF THE 3D SEGMENTATION ALGORITHM

- Given a 3D MRI initialize a 6-connected graph $G^{[0]} = (V^{[0]}, W^{[0]})$ where $V^{[0]}$ is the set of image pixels and $W^{[0]}$ is defined according to (6).
- Repeat recursively for $s = 1, 2, \ldots$ construct $G^{[s]}$ from $G^{[s-1]}$, as follows:
  1) Seed Selection: Select a representative set of nodes $V^{[s]}$, such that $V^{[s-1]} \setminus V^{[s]}$ is strongly connected to $V^{[s]}$.
  2) Define $P = P^{[s-1]}$ the interscale interpolation matrix (by Eq. 5).
  3) Calculate $W^{[s]} \approx P^T W^{[s-1]} P$ by weighted aggregation (Eq. 6).
  4) For each node $v \in V^{[s]}$ calculate aggregative features (Sec. II-B).
  5) Modify $W^{[s]}$ according to aggregative features (Eq. 9).

p and q in the coarse graph is assigned with the weight:

$$w_{kl}^{coarse} = \sum_{p \neq q} P_{pk} w_{pq} P_{ql}. \tag{6}$$

w_{pq}^{coarse} is also called the **coupling weight** between aggregates k and l. Intuitively, the coupling weight between a pair of coarse aggregates (left hand side of (6)) is the weighted sum of the coupling weights between their sub-aggregates (right hand side of (6)).

Denoting the scale by a superscript $G^{[s]} = (V^{[s]}, W^{[s]})$. Note that since $u^{[s-1]} \approx Pu^{[s]}$ the relation Eq. (3) inductively implies that a similar expression approximates Γ at all levels so that the saliency measure Γ can be written as

$$\Gamma = \frac{u^T L u}{\frac{1}{2} u^T W u} \approx \frac{U^T P^T L P U}{\frac{1}{2} U^T P^T W P U}. \tag{7}$$

However, $W^{[s]}$ is further modified to account for aggregative properties (Sec. II-B). We modify $w_{pq}^{[s]}$ between a pair of aggregates p and q at scale s by multiplying it with an exponentially decreasing function of their aggregative properties distance (see eq. 9).

Table I summarizes the segmentation algorithm.

*2) **Handling Anisotropic Data:*** The are many cases where the MRI data is anisotropic, less vertically resolved. The SWA algorithm, however, assumes that the voxels in the fine level are equally spaced, since the initial graph does not take into account the distances between neighbors (see Eq. 1). Ignoring this effect may lead to distorted segmentations. To solve this problem we modify the algorithms as follows. During the first few coarsening steps we consider each 2D slice separately in performing seed selection and inter-scale interpolation (steps 1-2 in Table I), allowing non-zero interpolation weights only between nodes of the same slice. The rest of the steps (steps 3-5 in Table I) are performed on the full 3D graph, i.e., taking into account inter-slice couplings. This procedure is repeated until the inner-slice and inter-slice distances are approximately equal. Subsequent coarsening steps consider the full 3D graph.

For example, consider data with 5_{mm} slice thickness versus $1_{mm} \times 1_{mm}$ in-slice resolution. Every coarsening step of the

SWA algorithm typically reduces the number of nodes by a factor of 2.5-3. Consequently, if we apply the algorithm to a 2D slice, the distance between neighboring nodes in a slice grows at every level by a $\sqrt{2.5}$-$\sqrt{3}$ factor on average, so three coarsening steps are needed to bring the inner- and inter-slice distances to be roughly equal.

*3) **Multi-channel Segmentation**:* A major aspect of MR imaging is the large variety of pulse sequences that can be applied. These sequences produce different images for the same tissue, highlighting different properties of the tissue. Common MR channels employed for MS lesion detection include fluid attenuated inversion recovery (FLAIR), T1-, T2-, and proton density (PD)-weighted images. We incorporate multi-channel data in the algorithm in a fairly straightforward manner.

Generally, when dealing with multi-channel data we use the popular Statistical Parametric Mapping (SPM) software package [23] to perform a 3D affine transformation in order to align the different sequences provided for each subject. Thus in this work, The alignment between T1-weighted and Dual-Echo images was achieved in the acquisition phase, by acquiring the T1-weighted immediately after the Dual-Echo and using the same positioning parameters (recorded on the scanner). Hence, given the multi-channel aligned scans, each voxel now includes a vector of intensities. The initialization step (Eq. 1) is modified to determine the initial weights utilizing intensity information from all m channels as follows:

$$w_{ij} = \exp -(\sum_{c=1}^{m} (\alpha_c)^2 (I_i^c - I_j^c)^2)^{\frac{1}{2}} \tag{8}$$

where α_c are pre-determined constants ($\alpha_{T2} = 15, \alpha_{PD} = \alpha_{T1} = 10$) and I_i^c is the intensity of voxel i in channel c. The selection of parameters is based on our experience with MRI data, where in a single channel experiment $\alpha = 15$ is used. Here, in the multi-channel experiment α_{PD}, α_{T1} are lower, putting more emphais on T2 intensity contrasts effects in the segmentation process. Genereally, all parameters used in the experiments were determined based on two randomly chosen calibration scans for each type of data, which were not used later in the training and testing experiments (see Sec. III-B).

Following the weights initialization, we maintain different sets of aggregative features for every channel (see Sec. II-B below) and use these properties to modify the edge weights at coarser levels. As described in table I-step 5, the coarse graph couplings are modified by the separate channels aggregative properties and allow different coefficient weighting for each channel and property. Some aggregative properties are common to the entire *mixel* (multi-channel voxel) while others are characteristics of each channel's intensity properties. Let m,p denote the total number of channels and scales respectively, then the influence of the multiscale average intensity and variance statistics on the coupling between two aggregates k,l can be considered by multiplying the coupling with the following term:

$$\exp -(\sum_{c=1}^{m} (\gamma_c)^2 (I_k - I_l)^2)^{\frac{1}{2}} \exp -(\sum_{c=1}^{m} (\beta_c)^2 (\Delta\nu_{kl})^2)^{\frac{1}{2}} \tag{9}$$

where γ_c and β_c are coefficient parameters that control the weight of the different measures in the different c channels, and

$\Delta\nu_{kl}$ refers to the average of variances (see Table II below) which is defined as:

$$\Delta\nu_{kl} \quad = \quad \frac{1}{p}\sum_{s=1}^{p}\left(\frac{2(\nu_k^{[s]}-\nu_l^{[s]})}{(\nu_k^{[s]}+\nu_l^{[s]})}\right)^2 \tag{10}$$

B. Feature Extraction

Lesions can often be characterized by properties of aggregates that emerge at intermediate scales, and are difficult to extract by any uni-scale procedure. Such properties may include, for instance, intensity homogeneity, principal direction of the lesion, and intensity contrast with respect to neighboring tissues. Voxel-by-voxel analysis is limited in its ability to utilize such scale-dependent properties.

We refer to such properties as *aggregative features*. The weighted-aggregation scheme provides a recursive mechanism for calculating such properties along with the segmentation process. Following interaction with expert radiologists we have selected a list of features relevant to the problem domain. We use these properties for two purposes. First, we use these aggregative properties to affect the construction of the segmentation pyramid (see Eq. 9). Second, these properties are available for the classification procedure below (Sec. II-C). The actual effect of each of these features however is determined in training by an automatic learning process. Table II lists the features for an aggregate k at scale s which are included in the aggregates **high dimensional feature vector**.

Construction of the classifier based on the features in Table II requires consideration of the inter-subject and intra-subject variability, therefore all features were normalized for each brain as described in the Table. Section II-C.2 will present the significant role of the different features.

1) ***Aggregative Features****:* For an aggregate k in scale s we express an aggregative property as a number reflecting the weighted average of some property q which is measured in a finer scale r, $(r \leq s)$. For example, the average intensity of k is an aggregative property, since it is the average over all intensities measured at the voxels (nodes of scale $r = 0$) that belong to k. More complex aggregative properties can be constructed by combining several properties (e.g., variance of average intensities below, which combines the average intensity and average squares of intensities of k) or by taking averages over aggregative properties of finer scales (e.g., average of variances below). We denote such a property by $Q_k^{[r][s]}$, and shorten this to $Q_k^{[r]}$ when the context is clear.

In addition to these properties we can define binary aggregative properties, reflecting relations between two aggregates k and l at scale s. Such properties, denoted by $Q_{kl}^{[s]}$, are useful for describing boundary relations between neighboring tissues, e.g., surface area of boundary between k and l or the contrast between the average intensity of an aggregate k and the average intensity of its neighbors. Figure 3 illustrates several features computed for the segments throughout the aggregation process.

TABLE II
AGGREGATIVE FEATURES INCLUDED IN THE HIGH-DIMENSIONAL FEATURE VECTOR f OF AN AGGREGATE k

Graph measures:

- **Saliency:** Γ (Eq. 3)

Intensity statistics:

- **Average intensity:** of voxels in aggregate k, denoted $\bar{I}_k^{[0]}$. Normalized by the average intensity of the intracranial cavity (IC) (see details in Sec. III-B).
- **Maximum/Minimum intensity:** $\mu_k^{[2][s]}$ maximal/minimal average intensity of the sub-aggregates at scale 2. Normalized by dividing in $\bar{I}_k^{[0]}$
- **Variance of average intensities of scale r:** $Var^{[r]} = \bar{I}^{2[r]} - (\bar{I}_k^{[0]})^2$, where $\bar{I}^{2[r]}$ denotes the average of $(\bar{I}_l^{[0][r]})^2$ for all sub-aggregates l of k at scale r. Normalized by $\bar{I}_k^{2[0]}$.
- **Average of variances:** of scale r denoted $\bar{\nu_k}^{[r]}$ where $\nu_k^{[r][r]} = Var^{[0][r]}$.
- **Average intensity Proportions:** Proportion between Average intensity of the different channels.

Shape:

- **Volume:** $m^{[0]}$ is the aggregate volume in voxel units. Normalized by the IC volume.
- **Location:** $\bar{x}^{[0]}$, $\bar{y}^{[0]}$, $\bar{z}^{[0]}$. Each scan is brought to a common coordinate system using the Statistical Parametric Mapping (SPM) software package [23], which registers a scan to an atlas composed of subjects average of 152 T1-weighted scans.
- **Shape moments:** The length, width, depth ($L^{[0]}$, $W^{[0]}$,$D^{[0]}$ respectively), and orientation are specified by applying principal component analysis to the covariance matrix of the aggregate. Normalized by the Corresponding values of shape moments measured for the entire IC.
- **Intensity moments:** averages of products of the intensity and the coordinates of voxels in aggregate k, denoted $\overline{Ix}^{[0]}$,$\overline{Iy}^{[0]}$,$\overline{Iz}^{[0]}$. The normalization is performed by the following expression

$$\frac{\overline{Ix}^{[0]} - \bar{I}^{[0]}\bar{x}^{[0]}}{(Var^{[r]})^{\frac{1}{2}}(\bar{x_k}^{2[0]} - (\bar{x}_k^{[0]})^2)^{\frac{1}{2}}} \tag{11}$$

- **Distance transform:** For each aggregate, assigns the Euclidean distance between its location and the nearest pixel out of the intracranial cavity (IC) divided by the maximal IC distance for that brain.

Neighborhood statistics:

- **Boundary surface area:** denoted B_{kl}. B_{kl} refers to the surface area of the common border of aggregates k and l. It is accumulated by weighted aggregation such that all the weights on the finest graph are set to 1. The boundary surface area was normalized by $(m^{[0]})^{\frac{2}{3}}$ where $m^{[0]}$ is the volume of aggregate k in voxel units.
- **Neighborhood Contrast:** defined as the difference between the average intensity of a segment and its **neighborhood average intensity**, formulated as:

$$< Constrast >_k = \bar{I_k}^{[0]} - \frac{\sum_l B_{kl}\bar{I_l}^{[0]}}{\sum_l B_{kl}} \tag{12}$$

Atlas statistics:

- **Tissue probabilities:** the average likelihood of finding gray matter(GM), white matter(WM), cerebro-spinal fluid (CSF) or the cerebellum (CE), in an aggregate k, denoted by $\overline{P}_{WM}^{[0]}$, $\overline{P}_{GM}^{[0]}$, $\overline{P}_{CSF}^{[0]}$, $\overline{P}_{CE}^{[0]}$. Computed using the SPM software [23] to align the subject's data and ICBM atlas probability maps [24], which represent the probability of finding a tissue type at a specified position.

*2) **Recursive Accumulation of Features**:* Aggregative properties of an aggregate k are in fact averages over its sub-aggregates. Such properties can be accumulated from one level of scale to the next with the interpolation weights determining the relative weight of every sub-aggregate. Below we describe how such properties are accumulated.

Consider a property q defined at some scale r. Let k be some aggregate at scale s ($r < s$), and suppose we wish to compute the average value of q in k, denoted $\bar{Q}_k^{[r][s]}$. Let j be an index over the aggregates of level r ($1 \leq j \leq |V^{[r]}|$), let $q_j^{[r]}$ denote the property q of j, and denote by p_{jk} the interpolation weights relating an aggregate j at level r with the parent aggregate k at level s. (Note that r and s may not be successive levels, and so we generalize interpolation weights to non-successive levels by defining the interpolation matrix $P^{[r][s]}$ as a product of interpolations $P^{[r][s]} = \prod_{t=r}^{s-1} P^{[t]}$. p_{jk} then is an element of $P^{[r][s]}$.) Then,

$$\bar{Q}_k^{[r][s]} = \frac{\sum_j p_{jk}^{[r][s]} q_j^{[r]}}{\sum_j p_{jk}^{[r][s]}} \tag{13}$$

To recursively compute this measure we accumulate the numerator and denominator separately, using the following formulas. Denote by $\mathbf{q}^{[r][r]} \overset{def}{=} (q_1^{[r]}, ..., q_{|V^{[r]}|}^{[r]})$ and by $\mathbf{m}^{[r][r]} \overset{def}{=} \vec{1} = (1, \ldots, 1)$ where $\mathbf{m}^{[r][r]}$ is of length $|V^{[r]}|$. Then, $\mathbf{q}^{[r][s]} \overset{def}{=} \mathbf{q}^{[r][s-1]} P^{[s-1]}$, and $\mathbf{m}^{[r][s]} \overset{def}{=} \mathbf{m}^{[r][s-1]} P^{[s-1]}$, $(r < s)$. Note that $\mathbf{m}_k^{[r][s]}$ is the number of sub-aggregates at scale r that compose the aggregate k at scale s. In particular, $m_k^{[0][s]}$, which is the number of voxels composing the aggregate k at scale s provides its volume. From these recursive relations we then calculate the required average

$$\bar{Q}_k^{[r][s]} = \frac{q_k^{[r][s]}}{m_k^{[r][s]}} \tag{14}$$

In this way the aggregative properties at each level s are calculated from information already accumulated at the immediately preceding level $(s-1)$. This computation is performed with the maintaining the linear complexity of the segmentation process.

To summarize, by the end of the segmentation procedure each aggregate is characterized by a **high-dimensional feature vector** f where each measurment described in Table II is a component in the vector.

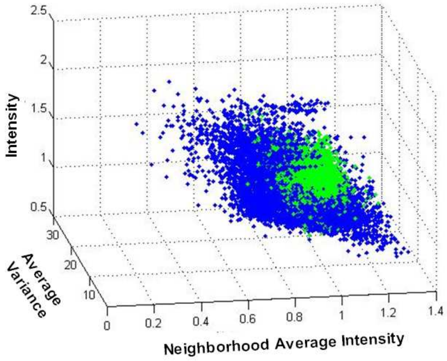


Fig. 3. 3D feature space. Neighborhood average intensity (x), average variance $\bar{\nu}_k^{[s][s]}$ (y) and average intensity (z) values presented for lesion (green) and non-lesion (blue) candidate segments.

C. Classification

Once an MRI scan is segmented and features are computed, each aggregate is characterized by a high-dimensional feature vector f (see Table II), and the classification stage can proceed. A classifier utilizing multiple decision trees ([25],[26]) is trained based on the aggregative features using data labeled by experts. Then, given an unlabeled scan the classifier is used to detect the lesions. Decision trees were used in the past by Kamber et al. [2] for MS segmentation. However here, we use multiple decision trees, with Fisher Linear Discriminant (FLD) [15] combined with a segmentation approach and a much richer set of features. Below we describe the the training and testing of segments (Sec. II-C.1) and how we use the classification results to determine the classification of individual voxels (Sec. II-C.3).

1) ***Multiple Decision Trees****:* To construct the decision tree classifier, a learning process is applied using MRI scans with MS lesions delineated by experts. The process obtains two kinds of data. (1) A collection of M candidate segments, $Cand = \{f_1, \ldots, f_M\}$, each is described by a high-dimensional feature vector f (each feature is normalized to have zero mean and unit variance), and (2) a mask indicating the voxels marked as lesions by an expert. Since the candidate segments may contain a mixed collection of lesion and non-lesion voxels we label as a lesion and denote by c_1, a segment in which $\geq 70\%$ of its voxels were marked by an expert as lesion. The threshold $\geq 70\%$ was selected so that it leads a training set, where the overlap of the c_1 class segments with lesions can allow them to be characterized by lesion properties. We further mark as non-lesions only those segments which do not contain lesion voxels at all and denote this class by c_2. The rest of the segments are ignored at the training stage.

We next use the training data to construct multiple decision trees. A subset of the candidate segments are randomly selected and used to construct a tree from the root downwards. At the root node all the labeled segments are considered and are repeatedly split into two subsets. At each tree node we apply a Fisher Linear Discriminant (FLD) [15] to the data determining the optimal separation direction $\overrightarrow{\phi}$ and threshold ξ that leads to a maximal impurity decrease. Figure 4 illustrates the recursive splitting process performed in each single decision tree.

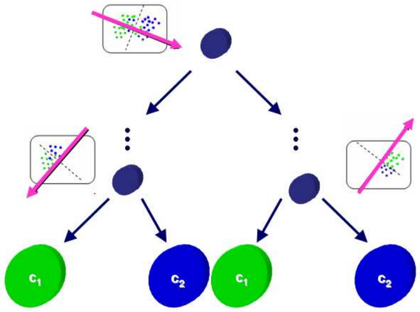


Fig. 4. A single classification tree. At each level of the tree we compute a two class Fisher Linear Discriminant (FLD) and find along the FLD direction the threshold split that achieves maximal impurity decrease based on entropy.

The impurity of the data at a certain node t is defined as:

$$E(t) = -\sum_{j=1}^{j=2} P(f \in c_j|t) log_2 P(f \in c_j|t) \tag{15}$$

where $P(f \in c_j|t)$ is the fraction of the training patterns f at node t in the tree that belong to class c_j. Given a splitting threshold ξ on $\overrightarrow{\phi}$, we denote the fraction of data points at t that goes to t_l and t_r with $P_l(\xi)$ and $P_r(\xi)$ respectively. We are looking for the optimal threshold ξ^* that maximizes the impurity decrease, i.e. $\xi^* = argmax_\xi \Delta E(t, \xi)$, defined as,

$$\Delta E(t, \xi) = E(t) - P_l(\xi)E(t_l) - P_r(\xi)E(t_r) \tag{16}$$

where $E(t_l)$ and $E(t_r)$ are calculated with respect to ξ. This training procedure results in a **forest** of K decision trees $T_1, \ldots, T_K$ each trained with a random selection of segments.

During the **testing phase** an unseen MRI scan is obtained. After segmentation and feature extraction we classify every high-dimensional feature vector f of a candidate segment by each of the K trees. Each tree T_q then determines a probability measure $P_{T_q}(f \in c_j)$ according to the distribution of training patterns in the terminal leaf node reached. These measures are integrated by taking their mean

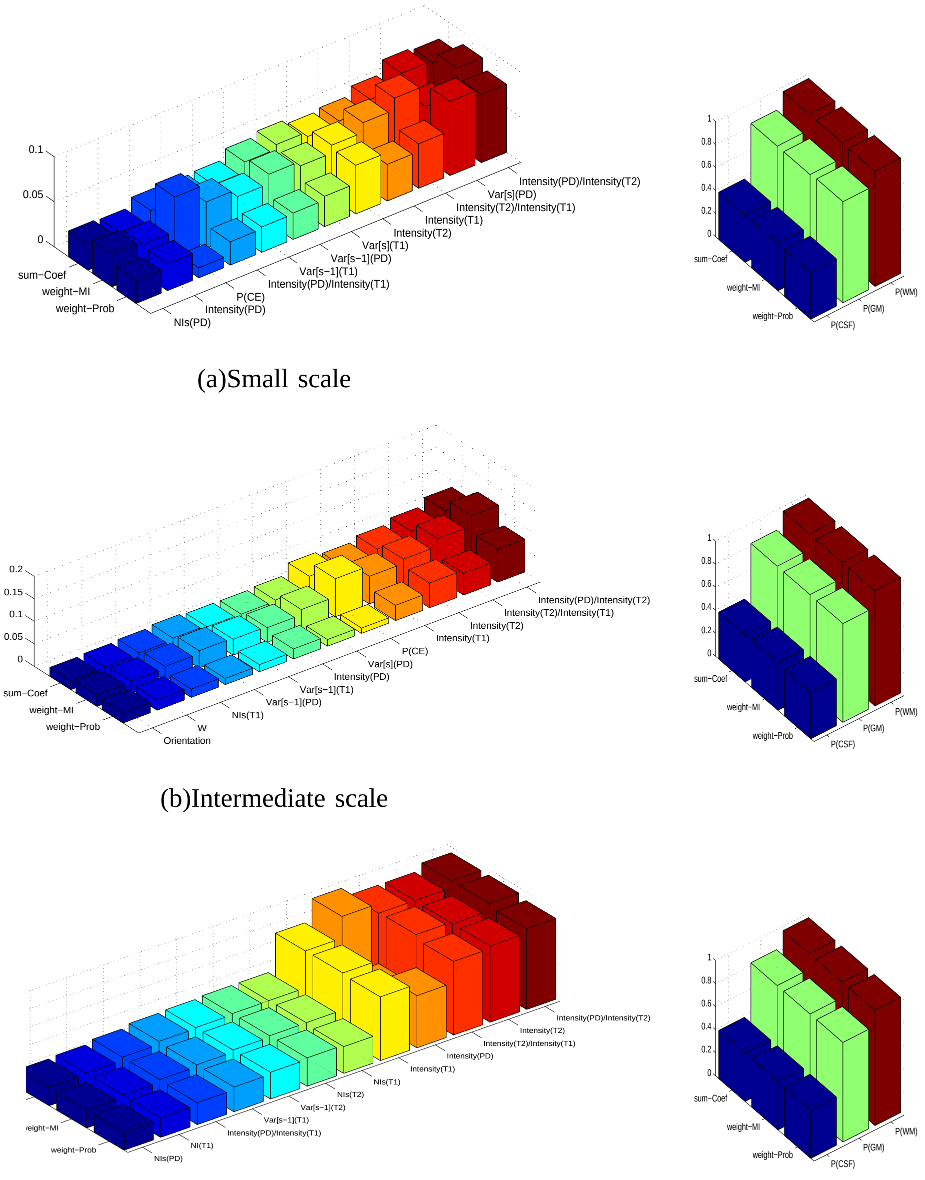
$$\frac{1}{K} \sum_{q=1}^{K} P_{T_q}(f \in c_j). \tag{17}$$

Finally, a test segment is assigned with the label c_j that maximizes this mean.

2) ***Significance of multiscale features:*** Figure 5,6 demonstrate the role of the different features in the classifer for both the multi-channel and Flair experiment respectively. Each figure presents the features ordered by their significance in the multiple decision trees from left to right. As described in section II-C.1 at each node in the tree, an optimal separating direction $\overrightarrow{\phi}$ is computed, which includes coefficients for each one of the features. Thus, the features significance in the multiple decision trees can be measured by computing three different summations:

- *sum-coef:*The absolute value of the feature coefficients used in all the tree nodes (ignoring the role of the node in the tree).
- *weight-NodeProb:* After weighting the coefficients of the feature vector by the node probability (proportion of training points at the node to the root).
- *weight-NodeMI:* After weighting the coefficients of the feature vector by the mutual information obtained by the nodes split (Eq. 16).

The figures show that the atlas probabilities are extremely significant in the classification process as expected. Other commonly used features in MS segmentation such as the average intensity and variance of average intensities also obtain high rank values, implying that the features selected automatically are reasonable. Still, the features used by the classifier are not voxel properties but aggregative properties. Moreover, the figures demonstrate that novel multiscale regional features such as measurements of the contrast to the neighborhood, or shape properties (e.g., width, length, orientation) can contribute to the classification process. The fact that these features are important characteristics of MS lesions is consistent with a large body of research on MS ([27]. The results, presenting a different set of features for the single and multi-channel experiments, and for the different scales, also show that different features are necessary for different recognition tasks. Our study can provide a useful tool



(a)Small scale

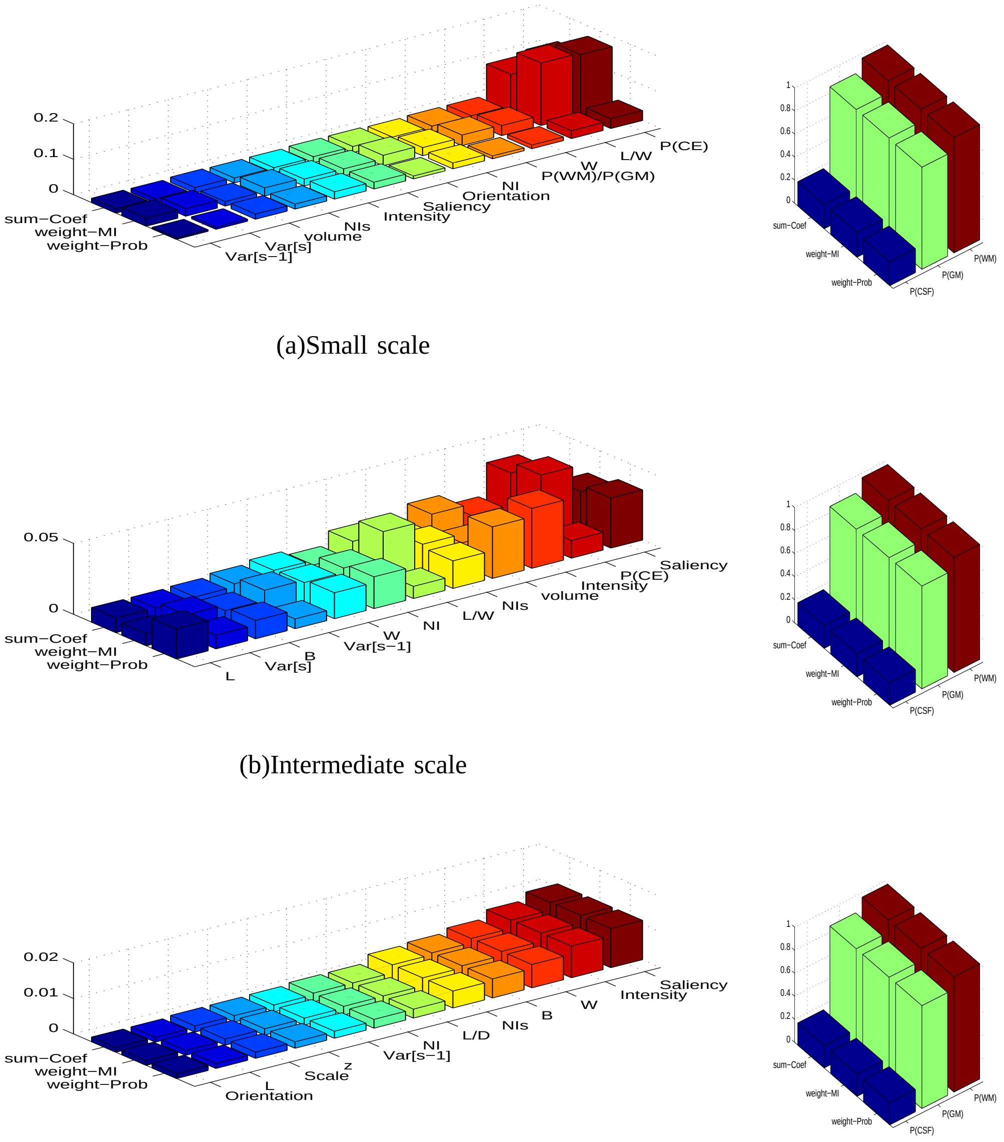
(b)Intermediate scale

(c) large scale

Fig. 5. Significance of Features in multi-channel experiment. Ordering the 20 most significant ordered from left to right. On the left 17 features ordered from low to high significance On the right a the 3 most significant features

for evaluating the importance of various features in a segmentation task. Moreover, we suggests that using a large bank of features and allowing an algorithm to determine the significant features automatically can be useful for many medical imaging applications assisting diagnosis.

*3) **Classification of Voxels**:* The output of the segmentation process includes the set of aggregates of the graphs in all scales and the interpolation weights between them. The classification process is applied to three sets of segmentation scales, *small, intermediate, and large* segments corresponding to scales $2, 3, \geq 4$ in the graph pyramid respectively. This separation was based on the idea that small, intermediate and large lesions share different attirbutes. Therefore, for each of these scales we construct a separate forest consisting of $K = 50$ trees, trained with a random selection of N_s patterns. In each experiment



(a)Small scale

(b)Intermediate scale

(c) large scale

Fig. 6. Significance of Features in FLAIR experiment. Ordering the 20 most significant ordered from left to right. On the left 17 features ordered from low to high significance On the right a the 3 most significant features

the size of the random subset selection is determined by 75% of the class size and twice of this amount for the Non-Class since there were many more Non-Class aggregates).

To measure the total lesion load (TLL) it is necessary to generate a result in terms of voxels. All candidates are projected onto the data voxels using the interpolation matrix. Therefore, the interpolation matrix (eq. 5) determines an association weight for each voxel and candidate. Voxels that belong to an segment k at some scale are defined as voxels whose interpolation weight to k is maximal. By the end of the classification, the classifier labels each candidate segments as lesion or non-lesion with some probability. The candidate segments in the different scales may overlap, so that a voxel may belong to more than one segment. The maximum probability (Eq. 17) over all candidates to which the voxel belongs, determines the probability of the voxel to be a lesion. The automatic segmentation results is based on the voxel's probability to be a lesion. Segmentations obtained with the automatic algorithm with varying values of probabilities psi were compared with the expert segmentations (see Fig. 9).

III. Results: Application to Multiple Sclerosis

Below we present validation results of employing our integrated system to two types of real MR data.

A. Complexity Analysis

The segmentation complexity is linear in the number of voxels with only several dozen of computer operations per voxel. The complexity for generating a tree classifier is

$$O(d^2 N_s \log(N_s) + d^3 N_s + dN_s(\log(N_s))^2) \tag{18}$$

where d is the number of features ($= 30$ for the single channel and 53 in the multi-channel due to the additional channels statistics) and N_s (≤ 15000) is the number of training patterns for one decision tree. The first term includes the $\#$ operations required to construct the FLD Generalized Eigenvalue problem, the second term includes the $\#$ of operations required for solving it, and the third term refers to the $\#$ of operations necessary for optimal splitting of the training points in each tree node. Therefore, the training complexity is dominated by $O(dN_s(\log(N_s))^2)$ and the testing complexity is $O(d\log(N_s))$ per one test sample.

In sum, the method is completely automated, with several parameters in the segmentation and classification components which are application dependant. Our segmentation implementation takes approximately less than 3,5 minutes per subject for the single-channel, multi-channel data respectively on a standard Xeon 1.7GHz PC. The training of the classifier which is done once in advance takes less than an hour for one multi-channel experiment and even less for the single-channel experiment. Where applying the classifier in the testing phase takes about 1 minute per subject.

B. Candidate Extraction

Before employing classification we apply several constraints to eliminate candidate segments whose properties differ considerably from those expected from a lesion. The same constraints were used for the both types of experiments (Sec. III-D,III-E on the multi-channel and FLAIR data correspondingly). Two randomly chosen calibration scans from each type of data were used to determine the parameters values and these scans were not used later in the training and testing experiments.

- We remove aggregates that include very dark regions. Average intensity < 1 and neighborhood contrast < -0.25 (Table II), on the channels used for manual tracing (PD and Flair in experiments III-D,III-E respectively). Since our database did not contain a sufficient amount of ”black T1 holes” examples to train on, at this stage of the algorithm we restricted ourselves to detection of active lesions only so that hypointense lesions are not detected. Dealing with ”black T1 holes” will require removing this constraint in the future.

- We include only aggregates included in the intracranial cavity (IC) and eliminate aggregates that overlap with anatomical structures where as a rule lesions do not develop (e.g. aggregates located in the eye area were removed). Binary masks for the IC were automatically generated based on registration of the ICBM atlas probability maps [24] to the subject's data using the SPM software [23]. Additionally, a mask for the eyes was generated based on the calibration scans. The eye area was traced in these scans and used to determine the coordinates of the eye area, these coordinates were later used to remove the eyes area from the IC mask in all scans.

C. Validation measures

Denote (S) as a set of voxels detected as lesions by our automatic segmentation and (R) as the set of voxels labeled as MS lesions in the 'ground truth' expert reference. We follow the definition of others [**?**] so that true positive(TP) voxels, are defined as S voxels ovelapping with expert outlines $|S \cap R|$. True negative (TN) voxels are defined as all IC voxels not outlined as lesions by experts $|IC \cap \bar{R}|$. False positive (FP) are those detected in S but not by R $|S \cap \bar{R}|$ and false negative (FN) are those identified by R but not by S $|\bar{S} \cap R|$. To evaluate the similarity between S and R, we use the **validation measures** listed in Table III which are commonly used in (e.g., [28],[7],[4],[10]). These measures are computed after obtaining the results in terms of voxels from all candidates detected as MS by the forest classifiers as described in sec. II-C.3. The measures are presented in Table V and Table VII for the multi-channel and FLAIR data respectively.

TABLE III
VALIDATION MEASURES

- **Sensitivity** S_e**:** True positive fraction $TP/(TP+FN)$
- **Specificity** S_p**:** True negative fraction $TN/(TP+FP)$
- **Accuracy** Ac**:** percentage agreement $(TN+TP)/(TN+TP+FN+FP)$
- **Dice similarity** κ **statistics:** $2|S \cap R|/(|S|+|R|)$
- **Correlation coefficient** R^2 correlation analysis between the total lesion load (TLL) measured as the number of voxels classified as MS lesion detected by R and S.

D. Validation on Real Multi-channel MR Data

To evaluate our method in comparison with other extensively validated studies in literature we performed a multi-channel experiment including a triplet of PD-, T2- and T1- weighted channels, which is similar to the type of data used in ([7], [4], [10]). The image data for the multi-channel experiment was acquired on a SIEMENS Magnetom Vision scanner (1.5T MR scanner). For each subject the data consists of a dual-echo sequence which is a Turbo Spin-Echo PD/T2-weighted image pair (TR=3300 ms; TE=16/98 ms; Echo Train Length=5) and a spin-echo T1-weighted image (TR=768 ms; TE=15 ms). Each channel contains 24 contiguous axial slices with a pixel size of $0.98_{mm} \times 0.98_{mm}$, slice thickness 5_{mm} (FOV=$250 \times 250_{mm}$; Matrix=256×256). The MR data used in both experiments was produced in the Scientific Institute Ospedale San Raffaele in Scientific Institute

TABLE IV
DETECTION RATES OBTAINED ON REAL MULTI CHANNEL DATA AVERAGED OVER TEN RANDOMIZED EXPERIMENTS.

Scale	lesion	non-lesion	Total
Small	0.887 ± 0.005	0.957 ± 0.005	0.955 ± 0.021
Interm	0.951 ± 0.003	0.975 ± 0.003	0.974 ± 0.018
Large	0.955 ± 0.003	0.987 ± 0.003	0.985 ± 0.014

TABLE V
CLASSIFICATION MEASURES FOR REAL MULTI CHANNEL MRI SETS, AVERAGED OVER TEN EXPERIMENTS.

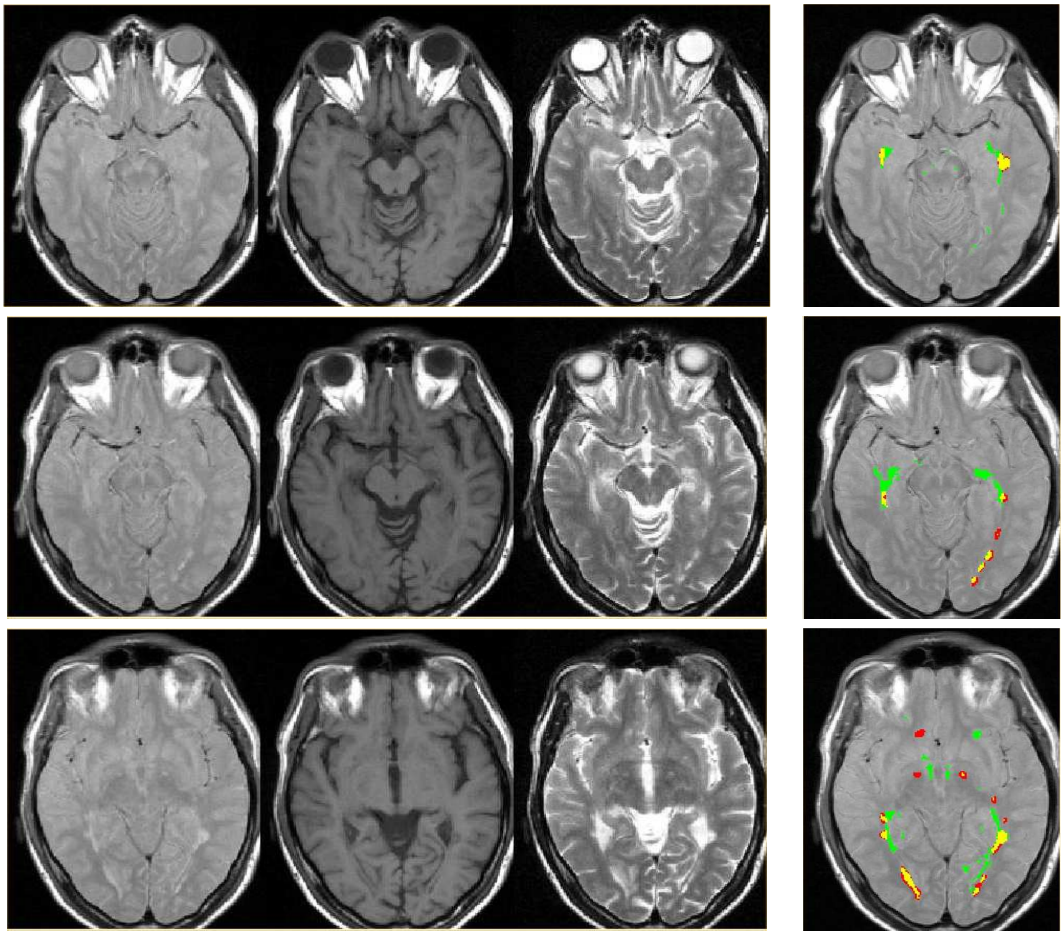
Probability (ψ)	Se	Sp	κ	Ac	R^2
0.5(none)	0.72 ± 0.11	0.95 ± 0.02	0.42 ± 0.09	0.94 ± 0.02	0.85
0.65	0.68 ± 0.11	0.96 ± 0.01	0.45 ± 0.09	0.95 ± 0.01	0.86
0.80	0.64 ± 0.12	0.97 ± 0.01	0.49 ± 0.09	0.96 ± 0.01	0.88
0.95(optimal κ)	0.55 ± 0.13	0.98 ± 0.01	0.53 ± 0.1	0.97 ± 0.01	0.89

Ospedale San Raffaele, Milan, Italy. The procedure for producing the lesion maps was the following: Two Neurologists by consensus identified hyperintense lesions on PD and Flair films in the multi- and single-channel experiments respectively. Using the marked films as reference, one trained Technincian outlined the contours of the lesions using a segmentation technique based on local thresholding [29]. The contours outlined from the Technician have been transformed in binary masks.

The tests were performed on MR data for 25 MS patients. Ten experiments were conducted. In each experiment, 75% of the patients were randomly selected for training. The test set consisted of the remaining patients of the multi-channel set. In each one of the ten experiments, three forests for the three different scales were generated. Table IV presents average detection rates for each scale over ten experiments. The **detection rate** measures the percentage of correct classifications of candidate segments in the test set. It is reported separately for the lesion class (c_1), non-lesion class (c_2) and total candidate set respectively (see Sec. II-C.1 for c_1,c_2 definitions). In addition, Table V will present results on terms of voxel rather than aggregates for all validation measures described.

Table V lists the average validation measures over the ten experiments for the multi-channel test in the entire brain. As described in section II-C.3 the automatic segmentation results is based on the voxel's probability to be a lesion. Therefore for each experiment we also assessed the scores behavior with varying values of probabilities (ψ). The results of S compared with the expert segmentations R are presented in figure 9(a). Table V presents several representative results where the rows correspond to $\psi = 0.5, 0.65, 0.8, 0.95$ respectively. The first row refers to the result where psi exceeds the value of 0.5, which includes the largest possible set of voxels detected as c_1. The last row refers the maximal κ point which as it appears in the results of both experiments the optimum was reached for $\psi \geq 0.95$. More motivation for using this graph will be given in the III-G.

Figures 7 and 8 demonstrate detection results for multi-channel real MRI data, obtained on inferior and superior slices respectively. Generally the results in the superior slices is better than the results in the inferior slices.



(a)Multi-channel data (PD-,T1-,T2-weighted)

(b)Detection

Fig. 7. Detection in inferior slices. Three inferior slices of multi-channel data, where each row corresponds to a different slice. From left to right: T1-, T2-, PD- weighted images and the area of the automatic segmentation (green) and 'ground truth' reference (red), where overlapping areas are colored in yellow.

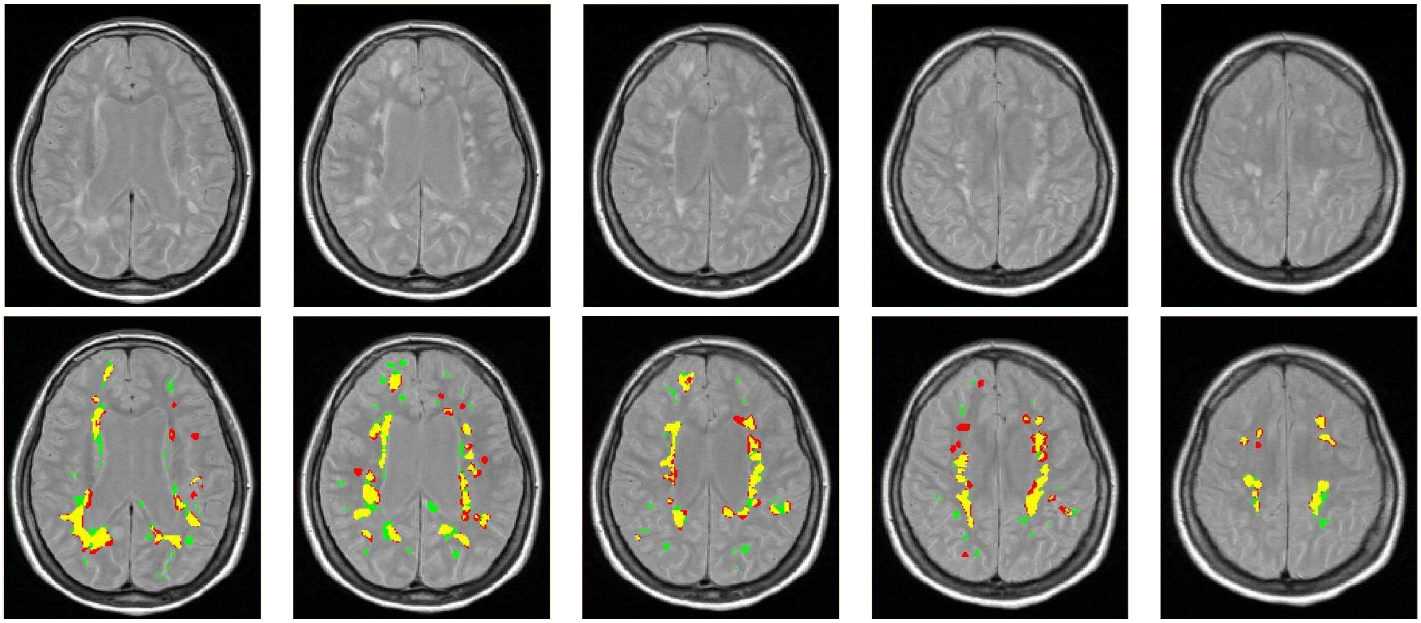


Fig. 8. Detection in superior slices. Top row: five superior slices of PD-weighted MRI. Bottom row: overlap of automatic segmentation (green) and 'ground truth' reference (red), where overlapping areas are colored in yellow.

TABLE VI
DETECTION RATES OBTAINED ON REAL FLAIR DATA AVERAGED OVER TEN RANDOMIZED EXPERIMENTS.

Scale	lesion	non-lesion	Total
Small	0.896 ± 0.006	0.974 ± 0.006	0.972 ± 0.014
Interm	0.943 ± 0.006	0.981 ± 0.006	0.98 ± 0.007
Large	0.973 ± 0.004	0.985 ± 0.004	0.984 ± 0.01

TABLE VII
CLASSIFICATION MEASURES FOR REAL FLAIR MRI SETS, AVERAGED OVER TEN EXPERIMENTS.

Probability (ψ)	Se	Sp	κ	Ac	R^2
0.5(None):	0.74 ± 0.1	0.96 ± 0.02	0.47 ± 0.07	0.96 ± 0.02	0.87
0.65:	0.71 ± 0.11	0.97 ± 0.02	0.50 ± 0.07	0.96 ± 0.02	0.87
0.80:	0.68 ± 0.11	0.98 ± 0.02	0.53 ± 0.07	0.97 ± 0.02	0.88
0.95(Optimal κ):	0.57 ± 0.14	0.99 ± 0.01	0.55 ± 0.09	0.98 ± 0.01	0.87

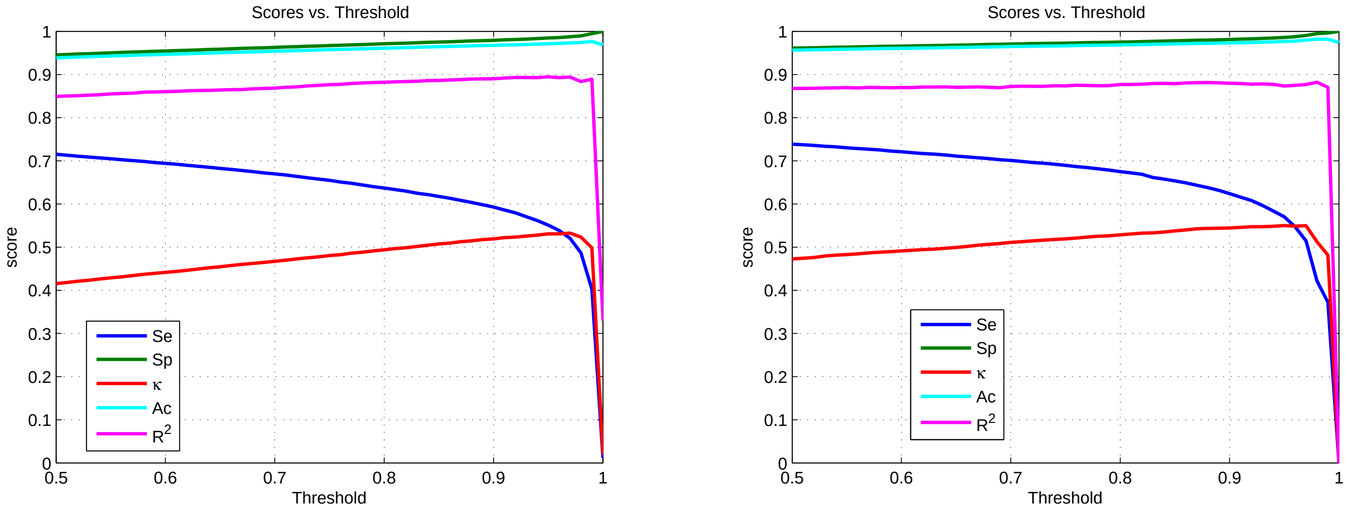
E. Validation on Real FLAIR MR Data

To evaluate the generalization ability of our approach, we tested our algorithm on real fluid attenuated inversion recovery (FLAIR) data [30]. In this case we used single channel FLAIR images since they are known for their high sensitivity to lesions, offering a diagnostic capability beyond other sequences. Brain MR images were acquired on a SIEMENS Magnetom Vision scanner (1.5T MR scanner). The FLAIR sequence used to acquire the images had the following parameters: TR=9500; TE=105; Inversion Time=2200; FOV=250×250 The acquisition was interleaved.

This study consists of 16 subjects for which MS lesions were manually traced by a human expert. The voxel size used is $0.97_{mm} \times 0.97_{mm}$ or $0.86_{mm} \times 0.86_{mm}$ (for 6 and 10 subjects respectively), with slice thickness 5_{mm} (24 slices). We divide the data as follows: set A includes examination of 12 patients and set B includes examinations of four additional patients which had a monthly follow up, so that four time points were available for each patient.

Throughout the classification stage ten experiments were conducted. In each experiment, nine patients from set A were randomly selected for training. The test set consists of the remaining patients of set A and all patients of set B. In each one of the ten experiments three forests for the three different scales were generated. Table VI presents average detection rates for each scale over ten experiments.In addition, Table V will present results on terms of voxel rather than aggregates for all validation measures described.

Table VII lists the average validation measures over the ten experiments, for test sets in the entire brain. For each experiment we also assessed the scores behavior with varying values of probabilities (ψ). The results of S compared with the expert segmentations R are presented in figure 9(b). Table VII presents several representative results where the rows correspond to $\psi = 0.5, 0.65, 0.8, 0.95$ respectively. The first row refers to the result where the voxel have probability above 0.5, which includes the largest possible set of voxels detected as c_1. The last row refers the maximal κ point which as it appears in the results of both experiments the optimum was reached for $\psi \geq 0.95$. More motivation for using this graph will be given in the III-G.



(a)

(b)

Fig. 9. Scores for multi-channel(a) and FLAIR(right) data as function of intensity threshold on the result.

Figure 11 presents a 3D view of MS lesions detected in the experiments on real FLAIR, and real multi-channel MRI data.

F. Volume Precision Over Time

We analyzed four sets of FLAIR images that were acquired over four months (set B). These data sets obtained validation results which are similar to the ones described in Sec. III-E.Generally, tests for robustness of reproducibility analysis should be performed on data rescanned repeatedly from the same brain. Here since the interval between two scans was not short, the volume may also vary due to actual changes in patient pathology. However, following the measure presented in [31] we performed a serial analysis and computed the ratio of volume difference between our detection and the 'ground-truth' divided by the mean of the two measurements. The analysis was performed based on the segmentation S obtained at the optimal κ point found in the FLAIR experiment III-E. The average results over time for each of the four subjects were $(0.1 \pm 0.06, 0.15 \pm 0.08, 0.07 \pm 0.06, 0.17 \pm 0.1)$ respectively.

Figure 10, presents for each of the subjects the total lesion load (TLL) detected by the automatic segmentation and the 'ground truth' reference over four points in time. As shown in the graph, the algorithm does not always follow the direction of the change in TLL. However, computing the average slope of the 'ground-truth' reference $\frac{(R_{t+1}-R_t)}{mean(R)}$ over all four subjects, shows very little changes in TLL as expected during four months 0.08 ± 0.08 (mean ± S.D). The algorithm may not be sensitive enough to detect such small changes. Still, considering the high correlation coefficient rate $(R^2 = 0.96)$ between the S,R overtime, this point should be further investigated, with more than one expert labeling, a larger time scale and more subjects.

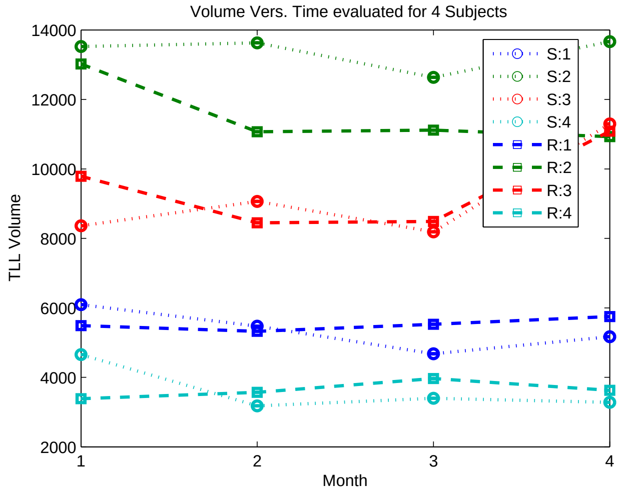
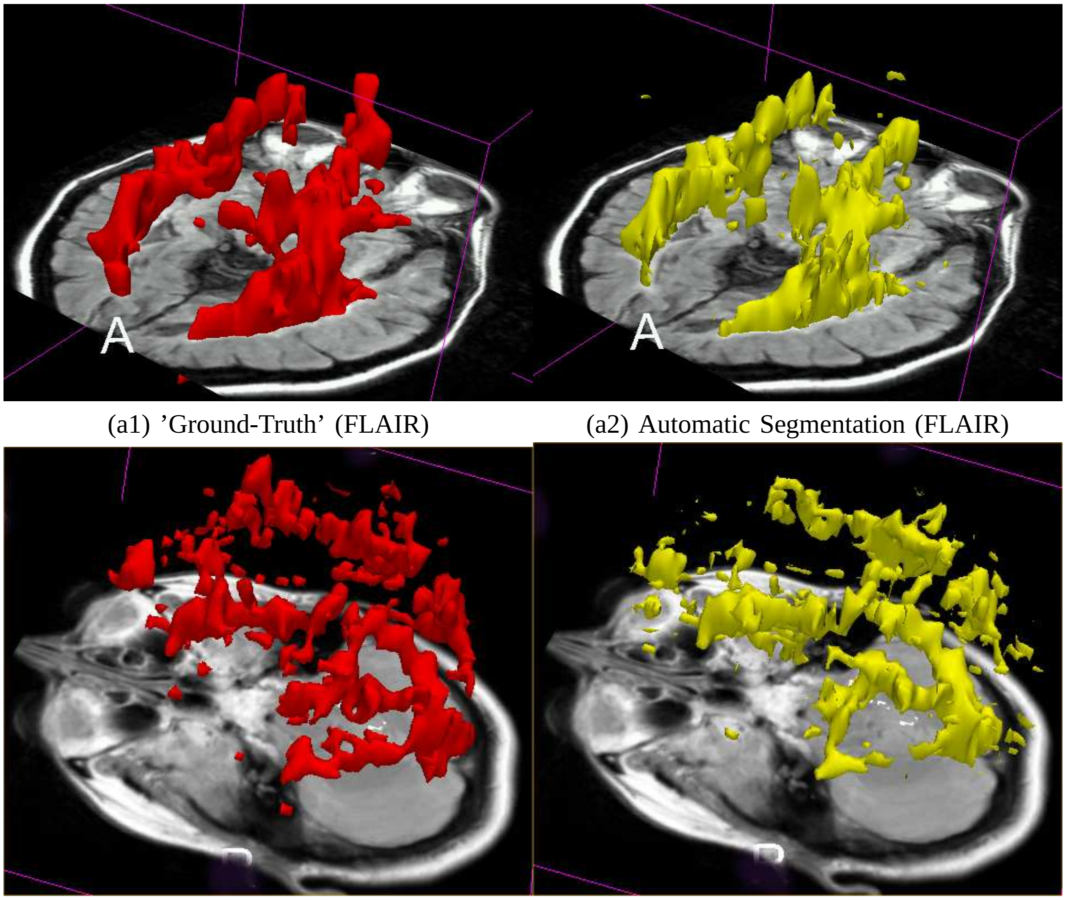


Fig. 10. Comparison of TLL volume obtained by S and R over time on four subjects, exemplified on set B (FLAIR).

G. Validation Analysis

Comparison to results reported in literature, demonstrates the difficulty of the MS detection problem and reveal potential obtained by our approach. To our best knowledge, studies reporting extensive validation results for automatic MS segmentation are performed on multi-channel data including T2-,PD- and T1-weighted images only. Therefore, we compare both our results on multi-channel and on FLAIR data to results reported on multi-channel data. The best correspondence results reported in [7] on multi-channel data were $\kappa = 0.45, 0.51$, for $5_{mm}, 3_{mm}$ slice thickness respectively with $R^2 = 0.96 - 0.98$. Where a similarity index of $\kappa = 0.58$ was found between two human experts. In [4] an average $\kappa = 0.6 \pm 0.07$ was obtained with $R^2 = 0.93$. Agreement between experts appears to fall in the same range, [4] report that the κ similarity between pairs of seven experts ranges from 0.51 to 0.67. Where the κ statistics in [4] were reported for a total of ten axial slice triplets of MRI scans. Recently resuls were published in [10] on several lesions subtypes. Their results were reported in terms of sensitivity $(70 - 75.2\%)$,specificity$(98.7 - 99.9\%)$, accuracy$(98.5 - 99.9\%)$,and $(R^2 = 0.96 - 0.98)$. The authors also report of correlation and agreement of lesion volume change over time $(R^2 = 0.715)$.

Since the previous papers either provide a κ score ([7],[4]) or a Sensitivity Specificity score [10] but not both, we decided to present the entire range for all measures in Figure 9. Evaluation of the results shows that comparing κ results at its optimal point $(\psi = 0.95)$ with papers that report κ values yields $\kappa = 0.53$,$\kappa = 0.55$ for the multi-channel and FLAIR experiments respectively. Which are higher than [7] but lower than [4] Comparison of the values reported by [10] at the basic $\psi = 0.5$ level, shows $Se = 0.72, Sp = 0.95, Ac = 0.94, R^2 = 0.85$ for the multi-channel and $Se = 0.74, Sp = 0.96, Ac = 0.96, R^2 = 0.87$ for the flair experiment respectively. Which are similar in sensitivity values and slightly lower in the specificity and accuracy values. Our correlation coefficients were lower compared to other studies, still, the correlations we obtained between manual and automatic measurements are all highly significant $(p < 0.0001)$ and our results for all spatial correspondence measures indicate that we obtain results not far from the state-of-the-art reports.



(a1) ’Ground-Truth’ (FLAIR) (a2) Automatic Segmentation (FLAIR)

(b1) ’Ground-Truth’ (multi-channel) (b2)Automatic Segmentation (multi-channel)

Fig. 11. 3D view of MS lesions detected by our automatic integrated segmentation and classification approach on real (a) FLAIR, and (b) multi-channel data . Comparison of expert labeling with automatic segmentation overlayed on an axial (a) FLAIR and (b) PD-weighted slice.

Inspection of the results shows that the results obtained by the real multi-channel (T1-,T2-,PD- weighted) are not as high as the results for the FLAIR experiment. We believe this is due to FLAIR’s higher sensitivity to MS lesions and also due to FLAIR’s higher specificity which allows avoiding many of the false positives (FP) detected in the multi-channel triplet. The higher specificity of FLAIR is a result of its ability to suppress the CSF signal, resulting in fewer FP in lesions near the ventricles and CSF containing sulci especially when the CSF is partially volumed with nearby brain parenchyma.

Although the algorithm is sensitive to the ψ parameter like [7], we consider reporting a result on a varying range, as an additional benefit of the algorithm. Since it is well known that the ’ground truth’ of the lesions may vary among different experts it might be helpful to provide a **soft classification** of the candidates rather than just a binary result.

IV. Discussion

We have presented a novel multiscale approach that combines segmentation with classification for detecting abnormal 3D brain structures. Our study focuses on analyzing 3D MRI brain data containing brain scans of Multiple Sclerosis patients. Our method is based on a combination of a powerful multiscale segmentation algorithm, a rich feature vocabulary describing the segments, and a decision tree-based classification of the segments. By combining segmentation and classification we are able to utilize integrative, regional properties that provide regional statistics of segments, characterize their overall shapes, and localize their boundaries.

We adapted the multiscale segmentation algorithm to handle 3D multi-channel MRI scans and anisotropic voxel resolutions. The rich set of features employed were selected in consultation with expert radiologists. All the features are computed as part of the segmentation process, and they are used in turn to further affect the segmentation process. The classification stage examines each aggregate occupied with its features and labels it as either lesion or non-lesion. This classification is integrated across scale to determine the voxel occupancy of the lesions. We have demonstrated the utility of our method through experiments on two types of real MRI data, including several modalities (T1, T2, PD and FLAIR). Comparison of our results to other automated segmentation methods applied to Multiple Sclerosis, yields similar κ and sensitivity values with lower specificity accuracy and correlation values. Therefore the detection rates obtained reflect the promise of our system.

Qualitatively evaluating the results we observe that the entire detected volume includes most of the lesion volume, while the main errors is caused by a high FP rate. This extra volume exhibited by FP volume should be further explored. In our experiments, the main extra volume usually surrounds the lesion volume and the number of voxels in the extra volume which are disconnected to any 'ground-truth' lesion divided by $|R|$ is significantly small compared to the FP rate. Preliminary assessment of our results indicates that this extra volume is somewhat related to other WM classes, e.g. 'dirty-appearing' WM (DAWM) [32]. Moreover, the delineation of lesion volume varies significantly between different experts, i.e, volume ratios reported in literature may exceed 1.5 ([7], [6], [4]). Therefore, since our data was labeled by a single rater, we may conclude that some of the FP volume is in the range of the inter-rater variability. Furthermore applying our approach on data with higher resolution (e.g. 3_{mm} slice thickness) may result in improved results as reported in [7].

Our approach is flexible with no restrictions on the MRI scan protocol, resolution, or orientation. Most automatic segmentation techniques focus on T2-, PD-, and T1-weighted data, whereas our work is designed to detect lesions using an unlimited set of sequences including FLAIR. In particular, FLAIR sequences are widely used clinically, yet to our knowledge there are only few reports on their automatic segmentation [33]. Unlike common approaches our method is not limited to finding the lesions in the WM only ([7],[**?**],[2],[5]), risking the omission of sub-cortical lesions. Furthermore, our learning process requires only a few training examples as shown specifically in the experiments.

We believe that our method can further be improved by better exploiting the rich information produced by the segmentation procedure. We plan to explore other features that can characterize lesions, as well as features that can characterize dirty appearing white matter (DAWM) and MS lesions subtypes. Also of importance is to incorporate more detailed prior knowledge of anatomic structures into the framework using a brain atlas. Preliminary work in this direction can be found in [34]. Finally, we wish to extend our approach and apply it to other tasks and modalities in medical imaging.

ACKNOWLEDGEMENT

Research was supported in part by the Binational Science foundation, Grant No. 2002/254, by the Israel Institute of Technology and by the European Commission Project IST-2002-506766 Aim Shape. Research was conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science.

REFERENCES

[1] A. Achiron, S. Gicquel, S. Miron, and M. Faibel, "Brain MRI lesion load quantification in Multiple Sclerosis: a comparison between automated multispectral and semi-automated thresholding computer-assisted techniques." *MRI*, vol. 177, pp. 85–106, 2001.

[2] M. Kamber, R. Shinghal, D. Collins, G. Francis, and A. Evans, "Model-based 3-d segmentation of Multiple Sclerosis lesions in mr brain images," *IEEE TMI*, vol. 14(3), pp. 442–453, 1995.

[3] J. Udupa, L. Wei, S. Samarasekera, Y. Miki, M. van Buchem, and R. Grossman, "Multiple Sclerosis lesion quantification using fuzzy-connectedness principles," *IEEE TMI*, vol. 16(5), pp. 598–609, 1997.

[4] A. Zijdenbos, R. Forghani, and A. Evans, "Automatic pipeline analysis of 3D MRI data for clinical trials: application to MS," *IEEE TMI*, vol. 21, pp. 1280–1291, 2002.

[5] S. Warfield, K. Zou, and W. M. Wells, "Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions," *J. of image guided surgery*, vol. 1(6), pp. 326–338, 1995.

[6] X. Wei, S. Warfield, K. Zou, Y. Wu, X. Li, A. Guimond, J. Mugler, R. Benson, L. Wolfson, H. Weiner, and C. Guttmann, "Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy." *JMRI*, vol. 15, pp. 203–209, 2002.

[7] K. Van-Leemput, F. Maes, D. Vandermeulen, A. Colcher, and P. Suetens, "Automated segmentation of Multiple Sclerosis by model outlier detection," *IEEE TMI*, vol. 20, pp. 677–688, 2001.

[8] W. M. Wells, W. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE TMI*, vol. 15, pp. 429–442, 1996.

[9] S. Warfield, A. Robatino, J. Dengler, and F. Jolesz, "Adaptive template moderated spatially varying statistical classification," *MIA*, vol. 4(1), pp. 43–55, 2000.

[10] Y. Wu, S. Warfield, I. Tan, W. W. 3rd, D. Meier, R. van Schijndel, F. Barkhof, and C. Guttmann, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel mri," *NeuroImage*, vol. 32(3), pp. 1205–15, 2006.

[11] R. Kikinis, C. R. G. Guttmann, D. Metcalf, W. M. W. III, G. J. Ettinger, H. L.Weiner, and F. A. Jolesz, "Quantitative follow-up of patients with multiple sclerosis using mri: Technical aspects," *J Magn Reson Imag*, vol. 9(4), p. 519530, 1999.

[12] C. G. A. C. G. Gerig, D. Welti and G. Szkely, "Exploring the discrimination power of the time domain for segmentation and characterization of lesions in serial mr data," *MICCAI*, p. 469480, 1998.

[13] D. Rey, G. Subsol, H. Delingette, and N. Ayache, "Automatic detection and segmentation of evolving processes in 3d medical images: Application to Multiple Sclerosis," *MIA*, vol. 6(4), pp. 163–179, 2002.

[14] A. Shahar and H. Greenspan, "A probabilistic framework for the detection and tracking in time of Multiple Sclerosis lesions," *IBSI*, 2004.

[15] J. R. Duda and P. Hart, Eds., *Pattern classification and scene analysis*. New York: John Wiley and Sons, 1973.

[16] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *CVPR*, 2003.

[17] E. Sharon, A. Brandt, and R. Basri, "Segmentation and boundary detection using multiscale intensity measurements," *CVPR*, pp. 469–476, 2001.

[18] M. Galun, E. Sharon, R. Basri, and A. Brandt, “Texture segmentation by multiscale aggregation of filter responses and shape elements,” *ICCV*, pp. 716–723, 2003.

[19] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt, “Hierarchy and adaptivity in segmenting visual scenes,” *Nature*.

[20] A. Akselrod-Ballin, M. Galun, J. M. Gomori, R.Basri, and A. Brandt, “Atlas guided identification of brain structures by combining 3d segmentation and svm,” *MICCAI*, 2006.

[21] A. Akselrod-Ballin, E. Eyal, M. Galun, E. Furman-Haran, J. M. Gomori, R. Basri, H. Degani, and A. Brandt, “Automatic 3d segmentation of the rat uterus in mri,” *SPIE*, 2006.

[22] A. Akselrod-Ballin, M. Galun, J. M. Gomori, M. Fillipi, P. Valsasina, R.Basri, and A. Brandt, “An integrated segmentation and classification approach applied to multiple sclerosis analysis,” *CVPR*, 2006.

[23] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *PAMI*, vol. 22, 2000.

[24] A. Brandt, S. McCormick, and J. Ruge, Eds., *Algebraic multigrid (AMG) for automatic multigrid solution with application to geodetic computations.* POB 1852, Fort Collins, Colorado: Inst. for Computational Studies, 1982.

[25] S. Frackowiak, K. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. Penny, Eds., *Human Brain Function*. Academic Press, 2003.

[26] J. Mazziotta, A. Toga, A. Evans, P. Fox, and J. Lancaster, “A probabilistic atlas of the human brain: theory and rationale,” *NeuroImage*, vol. 2, pp. 89–101, 1995.

[27] L. Breiman, J. Olshen, and C. Stone, Eds., *Classification And Regression Trees*. Wadsworth Belmont CA, 1984.

[28] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24(2), pp. 123–140, 1996.

[29] D. Goldberg-Zimring, H. Azhari, S. Miron, and A. Achiron, “3-d surface reconstruction of multiple sclerosis lesions using spherical harmonics,” *Magnetic Resonance in Medicine*, vol. 46, pp. 756–766, 2001.

[30] G. Gerig, M. Jomier, and M. Chakos, “Valmet: A new validation tool for assessing and improving 3D object segmentation,” *MICCAI*, pp. 516–523, 2001.

[31] M. Rovaris, M. Filippi, G. Calori, M. Rodegher, A. Campi, B. Colombo, and G. Comi, “Intra-observer reproducibility in measuring new putative markers of demyelination and axonal loss in multiple sclerosis: a comparison with conventional t2-weighted images,” *Journal of Neurology*, vol. 18, pp. 895–901, 1997.

[32] M. Rovaris, M. Rocca, T. Y. I. Yousry, B. Colombo, G. Comi, and M. Filippi, “Lesion load quantification on fast-flair, rapid acquisition relaxation-enhanced, and gradient spin echo brain mri scans from multiple sclerosis patients,” *MRI*, vol. 17(8), pp. 1105–10, 1999.

[33] C. Guttmann, R. Kikinis, M. Anderson, M. Jakab, S. Warfield, R. Kiliany, H. Weiner, and F. Jolesz, “Quantitive follow-up of patients with multiple-sclerosis using MRI: reproducibility,” *JMRI*, vol. 9, pp. 509–518, 1999.

[34] Y. Ge, R. Grossman, J. Babb, J. He, and L. Mannon, “Dirty-appearing white matter in Multiple Sclerosis: volumetric MRI and magnetization transfer ratio histogram analysis,” *AJNR Am J Neuroradiol*, vol. 24(10), pp. 1935–40, 2003.

[35] G. Dugas-Phocion, M. Gonzalez, C. Lebrun, S. Chanalet, C. Bensa, G. Malandain, and N. Ayache, “Hierarchical segmentation of Multiple Sclerosis lesions in multi-sequence MRI,” *IBSI*, 2004.

[36] B.Johnston, M. Atkins, B.Mackiewich, and M.Anderson, “Segmentation of Multiple Sclerosis lesions in intensity corrected multispectral MRI,” *IEEE TMI*, vol. 15(2), pp. 154–169, 1996.